Supplementary Information for: Attention-Based Functional-Group Coarse-Graining: A Deep Learning Framework for Molecular Prediction and Design

Ming Han,[1,2,*] Ge Sun,[1,3,4,5,*] Paul F. Nealey,[1,6] and Juan J. de Pablo[1,3,4,5,6,†]

[1] *Pritzker School of Molecular Engineering, University of Chicago, Chicago, IL, USA*
[2] *James Franck Institute, University of Chicago, Chicago, IL, USA*
[3] *Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, New York, NY, USA*
[4] *Department of Chemical and Biomolecular Engineering, Tandon School of Engineering, New York University, Brooklyn, NY, USA*
[5] *Department of Physics, New York University, New York, NY, USA*
[6] *Materials Science Division, Argonne National Laboratory, Lemont, IL, USA*

I. METHODS

A molecule is naturally a graph, where atoms are the nodes interconnected via chemical bonds. Compared to Simplified Molecular-Input Line-Entry System (SMILES) [1] strings or its variants, such atom graph provides a better representation of molecular topology. Furthermore, by recognizing the functional groups that are included in a molecule, it becomes possible to also represent it at a coarser level by a motif graph.

In the encoding phase, we first encode the local environment of individual atoms in a molecule by applying graph neural networks to the atom graph. The embedding vectors of atoms are then passed into the corresponding functional groups in the motif graph via a multi-layer perception. Subsequently, we can then encode the local environment of individual functional groups by applying graph neural networks to the motif graph. This process can be modeled through Eqs. (3-6) in the main text. Lastly, molecular embedding $\mathbf{h}^{\mathrm{m}}$ is derived based upon the embedding of the root motif $\mathbf{h}_0^{\mathrm{f}}$:

$$\mathbf{h}^{\mathrm{m}} = \boldsymbol{\mu}(\mathbf{h}_0^{\mathrm{f}}) + \boldsymbol{\sigma}(\mathbf{h}_0^{\mathrm{f}}) \circ \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{S1}$$

Instead of a deterministic mapping, we perform a statistical inference of $\mathbf{h}^{\mathrm{m}}$: we apply MLPs to derive the mean $\boldsymbol{\mu}$ and the standard deviation $\boldsymbol{\sigma}$, and sample the actual value of $\mathbf{h}^{\mathrm{m}}$ via the reparameterizing trick from a white noise $\boldsymbol{\epsilon}$. This variational design not only ensures the continuity of the latent space, but also yields an estimate on the posterior distribution $P(\mathbf{h}^{\mathrm{m}}|M) \approx Q(\mathbf{h}^{\mathrm{m}}|M) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$.

In the decoding phase, once the molecular embedding is obtained, the decoder reconstructs the molecule motif by motif, achieving high reconstruction fidelity. To further enhance structural consistency, we implemented a self-attention mechanism that guides the construction of the motif graph and the assignment of functional groups. This mechanism may offer advantages in more challenging settings.

Assuming that molecular embedding carries the key information of molecular structure, we progressively modeled this process conditional on the molecular embedding $\mathbf{h}^{\mathrm{m}}$, until completion using Eqs. (S2). The decoding process can be achieved through Eqs. (7-10) in the main text.

$$\begin{aligned} P(M) &= \int \mathrm{d}\mathbf{h}^{\mathrm{m}}\, P(\mathbf{h}^{\mathrm{m}})\, P\big(M \,\big|\, \mathbf{h}^{\mathrm{m}}\big) \\ &= \int \mathrm{d}\mathbf{h}^{\mathrm{m}}\, P(\mathbf{h}^{\mathrm{m}}) \prod_u P\Big(M_{\leq u} \,\Big|\, M_{\leq u-1},\, \mathbf{h}^{\mathrm{m}}\Big). \end{aligned} \tag{S2}$$

II. HIERARCHICAL GRAPH REPRESENTATION

A. Molecular embeddings

Our Hierarchical Graph Autoencoder demonstrates enhanced performance by using a functional-group-based motif vocabulary to create chemically meaningful embeddings. To validate this, we tested the same dataset of 6,000 adhesive

* These authors contributed equally to this work.
† Corresponding author. Email: jjd8110@nyu.edu

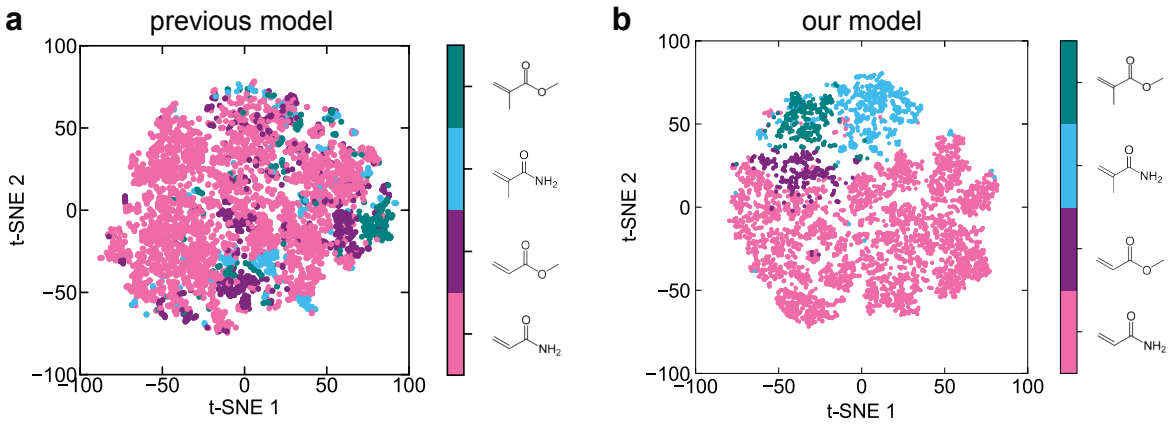

FIG. S1. **Comparison of model embeddings. a** Ineffective clustering in previous model embeddings. The approach relies solely on geometric data extraction of motifs, resulting in reduced chemical relevance and a lack of clear separations among monomer types. **b** Chemically meaningful embeddings in our autoencoder, identical to Fig. 2c in the main text for reference. We use t-SNE to map ten-dimensional embeddings into a two-dimensional space for both plots.

monomers with our model and a previous model. The previous model relies on geometrically derived motifs without incorporating prior domain knowledge, resulting in embeddings that lack chemical significance. Specifically, we use t-distributed stochastic neighbor embedding (t-SNE) to map high-dimensional latent space into a two-dimensional space for better visualization. The mapped embeddings fail to form distinct clusters and cannot distinguish different monomer types, as illustrated in Fig. S1a. To compare, Fig. S1b shows that our autoencoder effectively groups similar chemical structures into distinct clusters.

B. Reconstruction of molecules

To evaluate the efficacy of our hierarchical graph autoencoder, we compared its performance against established baselines: a string-based autoencoder [2], an atom-graph-based [3] autoencoder, and a data-driven motif extraction hierarchical variational autoencoder [4] (HierVAE). As shown in Table S1, our model reconstructs molecular structures with a success rate of approximately 95%, while the first two baseline methods struggled to exceed 50% under comparable conditions. Moreover, our method outperforms HierVAE by leveraging fewer than 100 well-established functional groups, which serve as a chemically meaningful vocabulary that faithfully captures the reactivity and structural diversity of polymeric and specialty chemicals. In contrast, purely data-driven motif extraction strategies often bias the model toward frequent ring motifs, fragmenting moiety into single/double bonds, thereby underserving chain-like and branched topologies and reducing overall fidelity.

Furthermore, our representation systematically integrates chemical information at multiple scales: (i) local atomic embeddings that capture fine-grained structural details, (ii) coarse-grained functional groups that encode chemically relevant building blocks, and (iii) a global molecular embedding that situates each molecule within the broader design space. This multi-level structure not only ensures robust reconstructions but also yields chemically interpretable latent features (see Fig. 2c), a marked improvement over existing data-driven hierarchical autoencoder models.

TABLE S1. **Comparison of methods**

Method	Reconstruction	
	Accuracy	Valid
Actual data	-	100%
SMILES VAE	29.7%	94.5%
Graph-based VAE	48.2%	100%
HierVAE	76.9%	100%
Our work	**95.3%**	100%

Reconstruction analysis and failure modes. In the acrylate monomer dataset (approximately 6,000 molecules), the average number of functional-group motifs per molecule is about seven. Fig. 2d shows representative successes: in each row, the reconstructed molecule is atom-bond isomorphic to the target, matching the motif inventory, ring presence and substitution pattern, branching locations, heteroatom placements, and chain lengths. These examples illustrate that the hierarchical decoder reliably preserves both global topology and local chemical detail. Empirically, reconstruction errors are dominated by *interface* predictions—i.e., selecting the atomic attachment site and the bond order when joining two correctly identified motifs—rather than by incorrect motif identity. The most frequent discrepancies involve symmetry-related attachment positions within a motif (e.g., alternative ring substitution indices). Once the motif graph and interfaces are fixed, full-atom decoding proceeds by motif templates and chemistry rules and contributes comparatively few errors. This points to attachment-site disambiguation and interface-aware conditioning as the primary levers for further improvement.

III. CHEMISTRY PREDICTION PERFORMANCE

A. Benchmark: QM9 dataset

To evaluate the predictive capability of our framework, we benchmarked it on the publicly available QM9 dataset [5]. The QM9 dataset is a widely used standard in computational chemistry and machine learning for quantum chemistry, containing approximately 130,000 small organic molecules with detailed quantum-chemical properties. It provides a comprehensive platform for assessing molecular representations and predictive models. Prior studies using QM9 have employed a range of machine learning techniques, including kernel regressors [6], graph neural networks [6, 7], and message-passing neural networks [8], which typically require large fractions of the dataset to achieve high accuracy.

In contrast, our model demonstrates strong predictive accuracy with limited data. We randomly selected 6,000 molecules for training and 2,000 for testing, corresponding to only about 5% of the full dataset. As shown in Fig. S2, the model achieves $R^2 \approx 0.97$–0.98 with low RMSE values in predicting the energies of the HOMO and LUMO, despite the small training size. This highlights the efficiency of our functional-group–based representation in capturing underlying chemical information.

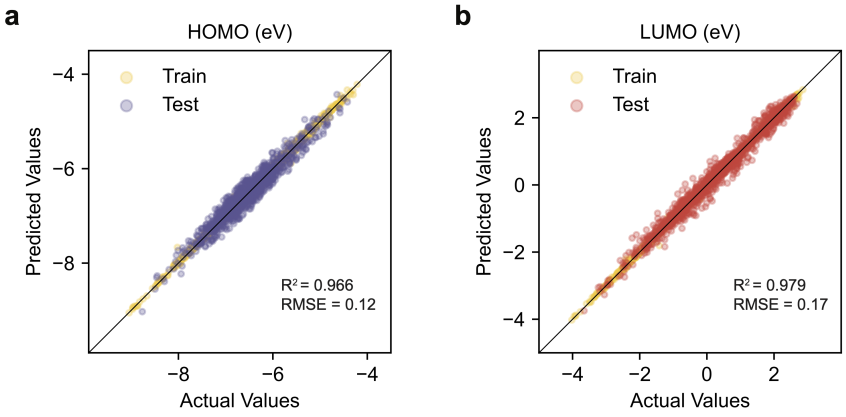

FIG. S2. **Prediction accuracy on the QM9 dataset: frontier orbital energies.** **a** Highest occupied molecular orbital (HOMO) energy and **b** lowest unoccupied molecular orbital (LUMO) energy predictions. Despite being trained on only 5% of the QM9 dataset, the model achieves $R^2 \approx 0.97$ with low RMSE values, demonstrating strong predictive performance under data-limited conditions.

To further assess generality, we extended the benchmarking to additional QM9 properties, including isotropic polarizability, electronic spatial extent, and heat capacity at 298 K. Results are shown in Fig. S3. Across all three properties, the model consistently achieves high predictive accuracy: $R^2 = 0.986$ for polarizability, $R^2 = 0.991$ for electronic spatial extent, and $R^2 = 0.978$ for heat capacity. These values are comparable to or better than reported results from atom-graph–based models trained on much larger fractions of the dataset. The ability of our model to perform well with limited training data can be attributed to its model architecture, which effectively captures the underlying chemical information of the molecules. This contrasts with traditional methods that rely heavily on large datasets to achieve similar levels of accuracy. Together, these results confirm that our hierarchical representation

provides both robustness and efficiency, enabling accurate predictions across diverse electronic and thermodynamic properties even under data-limited conditions.

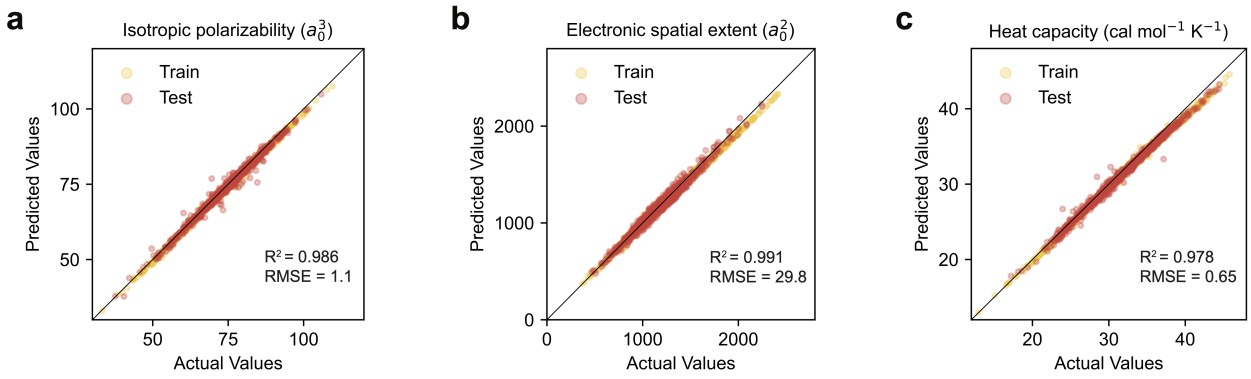

FIG. S3. **Prediction accuracy on the QM9 dataset: additional properties.** **a** Isotropic polarizability. **b** Electronic spatial extent. **c** Heat capacity at 298 K. Across diverse properties, the model achieves high accuracy, further confirming the robustness and generality of the functional-group–based representation.

B. Application: Monomers for adhesive polymeric materials

To test our model in a domain-specific setting, we evaluated a dataset of 600 monomer species relevant to adhesive polymeric materials. For each monomer, key thermophysical properties were obtained from all-atom molecular dynamics (MD) simulations, including cohesive energy (E_{coh}), heat of vaporization (ΔH_{vap}), isothermal compressibility (β), bulk density (ρ), radius of gyration (R_g), and glass transition temperature (T_g). The predictive performance of our full model is reported in Fig. 3b of the main text.

1. Ablation study

To clarify the contributions of different architectural components, we conducted two ablation studies using cohesive energy (E_{coh}) as a representative example.

First, we removed the *attention mechanism* from the property prediction module. As shown in Fig. S4, predictive accuracy decreased, with the test R^2 dropping from 0.93 (full model) to 0.85, and the RMSE increasing from 0.56×10^8 Pa to 0.67×10^8 Pa. This indicates that the attention layer plays a critical role in capturing higher-order interactions between motifs and in weighting chemically important groups more effectively. Without attention, the model still performs reasonably well, but the systematic loss of accuracy suggests that attention enhances the model's ability to focus on structural features most relevant to intermolecular cohesion.

Second, we excluded the *atom-level contributions* and relied only on motif-level features. As shown in Fig. S5, the performance decreased further ($R^2 = 0.82$, RMSE $= 0.07\times10^9$ Pa). This underscores the importance of fine-grained atomic detail, which provides critical information about local electronic environments and bonding patterns. While motif-level representations capture the overall scaffold of the molecule, atom-level descriptors refine the prediction by incorporating subtle effects such as heteroatom polarity, bond saturation, and substituent placement.

Taken together, these ablation studies confirm that both architectural elements—the atom-level contributions and the attention mechanism—are essential for optimal performance. The hierarchical design allows the model to leverage chemically meaningful motifs for efficiency, while atom-level corrections and attention weighting ensure predictive accuracy and interpretability. This balance is particularly valuable in data-limited, domain-specific applications such as adhesive monomer design, where both coarse-grained efficiency and fine-grained precision are needed.

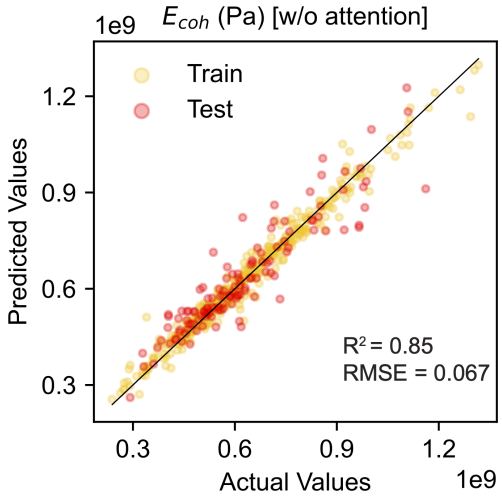

FIG. S4. **Ablation study: property prediction without attention.** Cohesive energy predictions (E_{coh}) for monomers when the attention mechanism is removed. The test performance decreases to $R^2 = 0.85$ with RMSE $= 0.67 \times 10^8$ Pa, indicating that attention improves accuracy by emphasizing chemically important motifs.

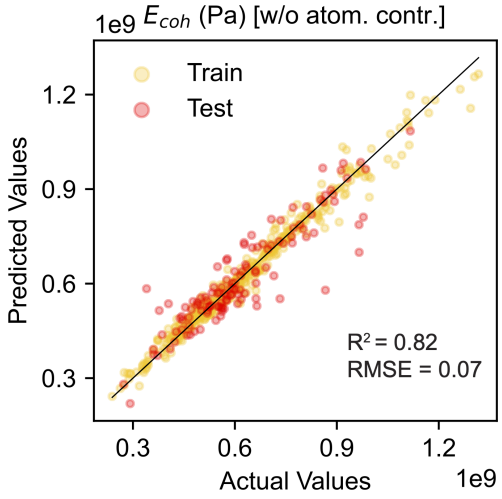

FIG. S5. **Ablation study: property prediction without atom-level contributions.** Cohesive energy predictions (E_{coh}) for monomers when only motif-level features are used. Test performance decreases further to $R^2 = 0.82$ with RMSE $= 0.07 \times 10^9$ Pa, underscoring the necessity of atom-level detail for accurate property prediction.

2. Multi-property Prediction

In practical molecular design, it is often desirable to optimize multiple properties simultaneously, since target performance metrics in polymeric materials are typically interdependent. To examine the capability of our framework in such settings, we trained the model to jointly predict both cohesive energy (E_{coh}) and glass transition temperature (T_g) using the adhesive monomer dataset.

The results, shown in Fig. S6, demonstrate that the model maintains high predictive accuracy even under this multi-task configuration. For E_{coh}, the joint model achieves $R^2 = 0.90$ with RMSE $= 0.06 \times 10^9$ Pa (Fig. S6a). For T_g, the performance reaches $R^2 = 0.84$ with RMSE $= 20.2$ K (Fig. S6b). While these values are slightly lower than the single-property models ($R^2 = 0.93$ for E_{coh} and $R^2 = 0.89$ for T_g), the decrease is modest and within expectation for multi-task learning, where the representation capacity must be shared across multiple targets.

From a physical perspective, this behavior reflects the fact that E_{coh} and T_g are governed by overlapping but distinct

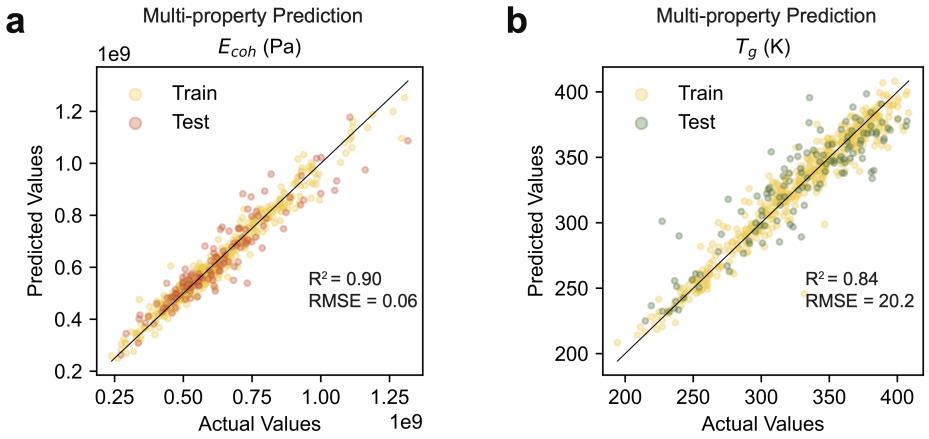

FIG. S6. **Multi-property prediction on adhesive monomer dataset. a** Joint prediction of cohesive energy density E_{coh}. **b** Joint prediction of glass transition temperature T_g. While performance is slightly lower than single-property models, the accuracy remains high, demonstrating the framework's capability for multi-objective prediction and design.

molecular determinants. Cohesive energy is primarily influenced by intermolecular interactions such as polarity and hydrogen bonding, whereas T_g depends additionally on collective phenomena such as chain rigidity and free volume. The ability of the model to learn both properties simultaneously, with only a minor reduction in accuracy, underscores the robustness of the hierarchical representation and its potential for practical multi-objective design.

These results suggest that our framework can naturally support multi-objective optimization strategies, for example by exploring Pareto fronts in latent space to identify candidate molecules that balance E_{coh} and T_g according to application-specific requirements.

3. Variability and Uncertainty Analysis

The accurate prediction of thermodynamic properties is critical for designing polymeric materials with targeted performance. However, some datasets are inherently more difficult to prepare and interpret. For example, measuring the glass transition temperature (T_g) in simulations requires systematic temperature scanning and equilibration at each step. Near the transition temperature, system dynamics slow down substantially, introducing additional uncertainty into the measurements.

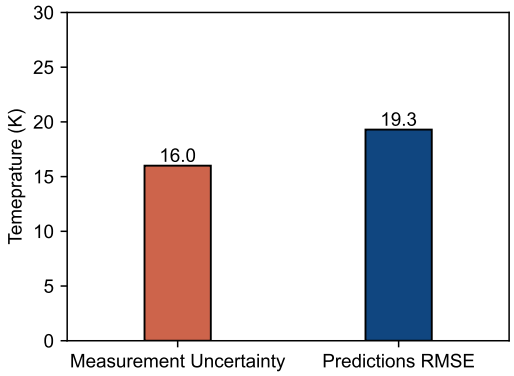

FIG. S7. **Measurement uncertainty and prediction RMSE for glass transition temperature (T_g).** The uncertainty from replicate simulations and the root mean square error (RMSE) of the model predictions are comparable, indicating that predictive performance is limited primarily by the quality of training data rather than the model itself.

In Fig. 3b of the main text, we showed that our model achieves high accuracy in predicting cohesive energy, heat

of vaporization, isothermal compressibility, bulk density, and radius of gyration. However, the predictive accuracy for T_g, while still significant, does not reach the same level as the other properties. Fig. S7 compares the measurement uncertainty from replicate simulations with the RMSE of our model predictions for T_g. The comparable magnitudes indicate that the primary limitation in T_g prediction stems from uncertainty in the training data itself, rather than deficiencies of the model. This underscores the intrinsic difficulty of generating reliable T_g data through simulations.

TABLE S2. **Performance across five random train/test splits.** Mean and standard deviation of test set R^2 values for T_g and E_{coh}.

Property	Mean R^2	Std. Dev.
Cohesive energy density (E_{coh})	0.93	0.01
Glass transition temperature (T_g)	0.89	0.02

To further evaluate statistical robustness, we repeated training with five independent random train/test splits and report the mean and standard deviation of performance metrics. Table S2 summarizes the results for T_g and E_{coh}. On average, the test R^2 is 0.89 ± 0.02 for T_g and 0.93 ± 0.01 for E_{coh}, demonstrating that the model's predictive performance is stable across different data partitions. Together, these analyses confirm that while T_g prediction is intrinsically more variable due to the underlying data, the framework remains robust and consistent across properties and training splits.

IV. MACHINE LEARNING PIPELINE FOR DOMAIN-SPECIFIC MOLECULAR DESIGN

In the main text, we demonstrated how the pipeline can generate novel acrylates with targeted glass transition temperatures and validated these predictions via molecular dynamics simulations. Here, we highlight a complementary application aimed at designing monomers with high cohesive energy density, further underscoring the versatility of our approach. Fig. S8a shows an overview of the workflow. We first generate 50,000 new acrylates that do not appear in the original dataset, then screen them using our integrated property predictor to identify the top 20 candidates. As illustrated in Fig. S8b, these newly proposed monomers display cohesive energies well above those in the initial database, with some values reaching up to five standard deviations from the mean. This result suggests that the pipeline can explore chemical regions beyond the training distribution, rather than merely reproducing near neighbors.

A closer look at the top designs reveals a common theme: many candidates feature polar moieties that foster strong intermolecular interactions, leading to higher cohesive energies. It is important to note, however, that focusing on a single property may yield molecules with limited synthetic viability or stability. Balancing multiple objectives would require expanding the current single-property optimization to a multi-property framework.

In either single- or multi-property settings, our pipeline functions as a first-pass screening layer that rapidly narrows the vast design space. The most promising candidates can then be evaluated using more computationally intensive methods or targeted experimental validation. By uniting generative exploration, high-throughput property prediction, and deeper validation steps, we establish a robust framework for machine-guided molecular design in polymer science and beyond.

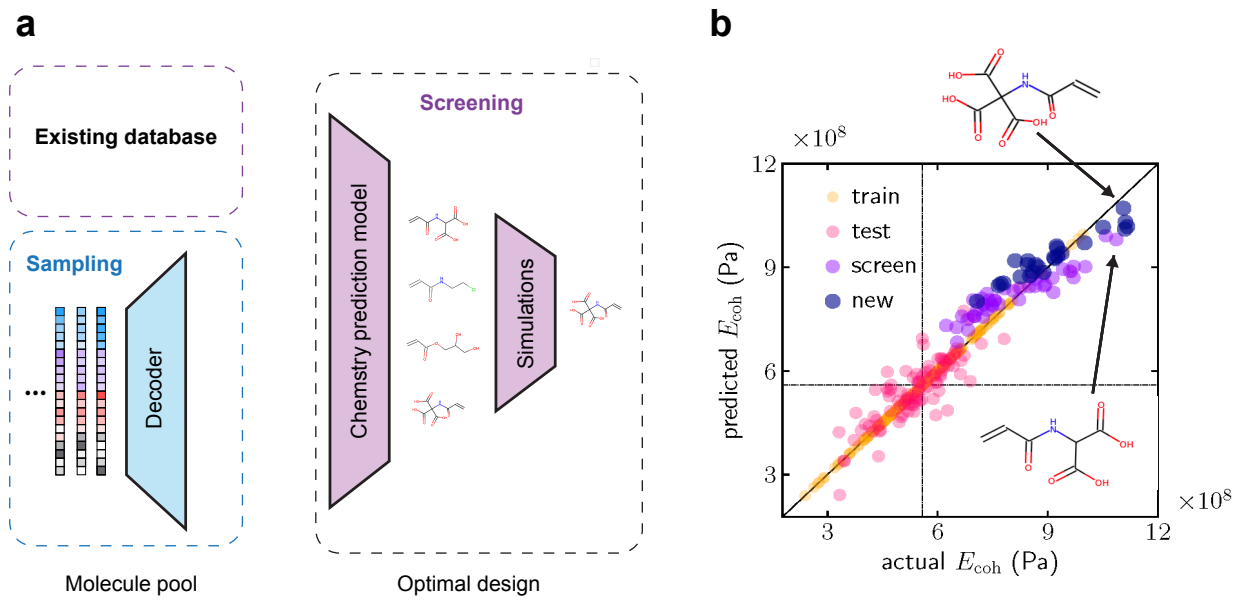

FIG. S8. **Automated machine learning pipeline for domain-specific molecular design (cohesive energy example).** **a** Schematic of the pipeline, where newly generated molecules are ranked by their predicted cohesive energy and selected for further validation. **b** The top candidates surpass the training-set average by up to 5σ, indicating the model's capacity to move beyond known design space.

SUPPLEMENTARY REFERENCES

[1] D. Weininger, Journal of chemical information and computer sciences **28**, 31 (1988).

[2] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, ACS central science **4**, 268 (2018).

[3] Q. Liu, M. Allamanis, M. Brockschmidt, and A. Gaunt, Advances in neural information processing systems **31** (2018).

[4] Jin, W., Barzilay, R. & Jaakkola, T. Hierarchical generation of molecular graphs using structural motifs. *International Conference On Machine Learning*. pp. 4839-4848 (2020)

[5] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, Scientific data **1**, 1 (2014).

[6] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, and O. A. von Lilienfeld, arXiv preprint arXiv:1702.05532 (2017).

[7] O. Rahaman and A. Gagliardi, Journal of Chemical Information and Modeling **60**, 5971 (2020).

[8] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, in *International conference on machine learning* (PMLR, 2017) pp. 1263–1272
.